Lecture 9:
Unsupervised learning

Nando de Freitas

www.cs.ubc.ca/~nando/340-2008/

September 2008

Outline

In the absence of labels ***y***, there are many useful patterns and structures that we can still find in the data. For example, we might be interested in:

• Novelty detection (e.g. detecting new strands of HIV).

• Data association (e.g. machine translation, multi-target tracking, object recognition from annotated images).

• Clustering (grouping similar items together).

In this lecture, we will introduce two of the most popular algorithms in machine learning and data mining: K-means and EM.

Clustering

K-means

K-means algorithm

1. **Initialisation:** Choose $k = 2$ means $\mu_{1:2}$ at random.
2. **Compute distances:** For $c = 1, \ldots, k$ and $i = 1, \ldots, n$ compute the distance $\|x_i - \mu_c\|^2$.
3. **Assign data to nearest mean:** To keep track of assignments, introduce the indicator variable z_i, such that

$$\mathbb{I}_c(z_i) = \begin{cases} 1 & \text{if } c = \arg\min_{c'} \ \|x_i - \mu_{c'}\|^2 \\ 0 & \text{otherwise} \end{cases}$$

   That is, $\mathbb{I}_2(z_i) = 1$ if observation x_i is closer to cluster 2. $\mathbb{I}_c(z_i)$ end up being the entries of an $n \times k$ matrix with only one 1 per row and many zeros.

K-means algorithm (continued)

4. **Update means:**

$$\mu_c = \frac{\sum_{i=1}^{n} \mathbb{I}_c(z_i) x_i}{\sum_{i=1}^{n} \mathbb{I}_c(z_i)}$$

5. **Repeat:** Go back to step 2. until the means and assignments stop changing.

Hard Vs Soft assignments

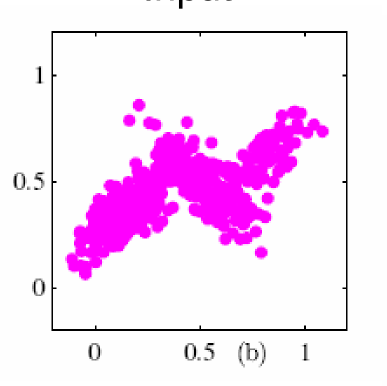
The problem with this algorithm is that the assignments are hard. Something is either this or that. Sometimes, however, we would like to say that something is this with probability 0.7 or that with probability 0.3.

We would like to find not only the means, but also the variances of each cluster and the probabilities of belonging to each cluster.

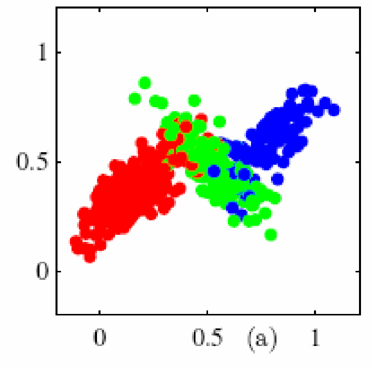
Clustering

Desired output

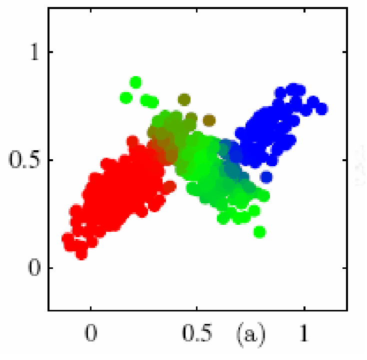
Input



Hard labeling

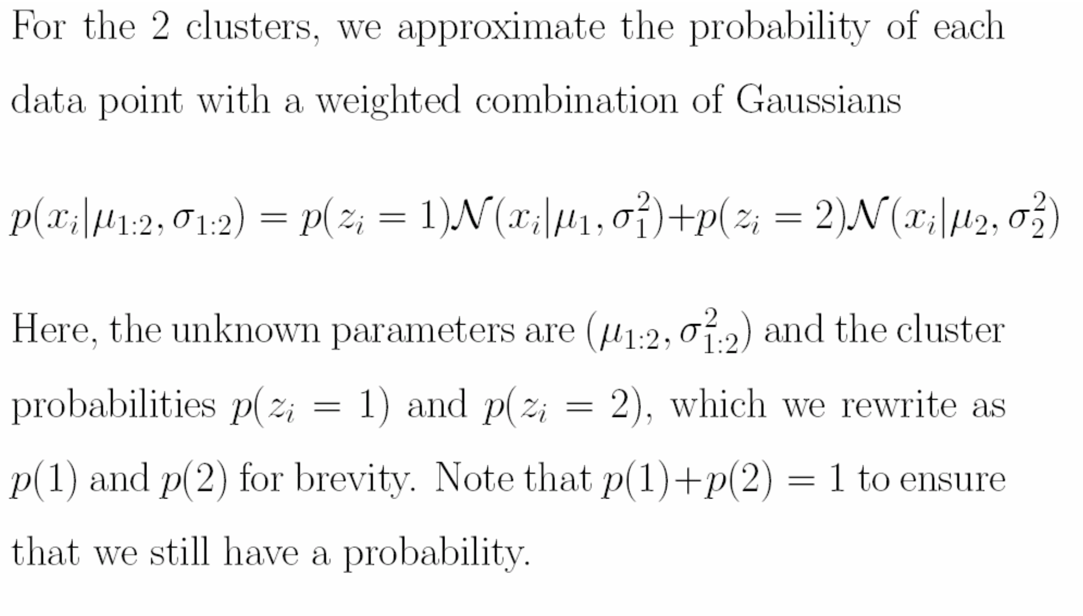


Soft labeling



K=3 is the number of clusters, here chosen by hand

Probabilistic approach



For the 2 clusters, we approximate the probability of each data point with a weighted combination of Gaussians

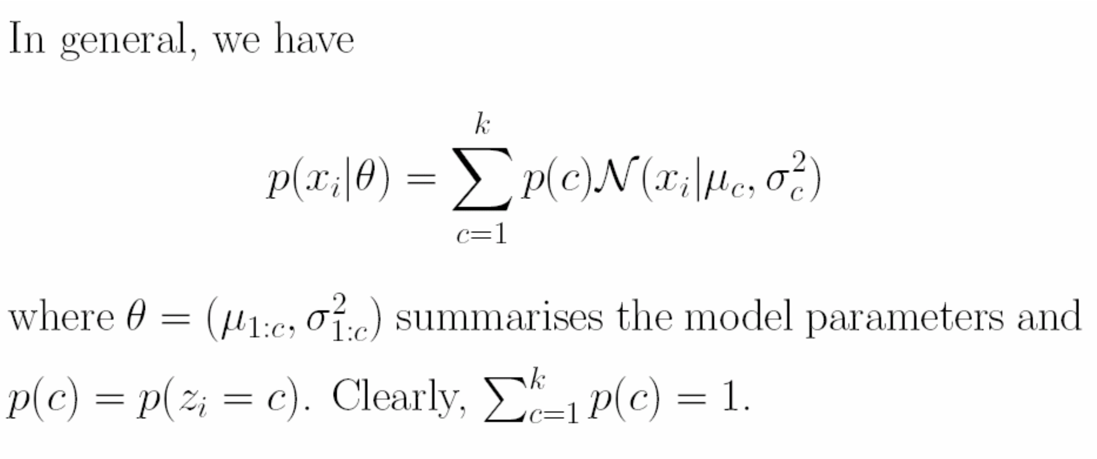
$$p(x_i|\mu_{1:2},\sigma_{1:2}) = p(z_i = 1)\mathcal{N}(x_i|\mu_1,\sigma_1^2)+p(z_i = 2)\mathcal{N}(x_i|\mu_2,\sigma_2^2)$$

Here, the unknown parameters are $(\mu_{1:2},\sigma_{1:2}^2)$ and the cluster probabilities $p(z_i = 1)$ and $p(z_i = 2)$, which we rewrite as $p(1)$ and $p(2)$ for brevity. Note that $p(1)+p(2) = 1$ to ensure that we still have a probability.

Probabilistic approach in 2D

Probabilistic approach in 1D

Probabilistic approach



In general, we have

$$p(x_i|\theta) = \sum_{c=1}^{k} p(c)\mathcal{N}(x_i|\mu_c, \sigma_c^2)$$

where $\theta = (\mu_{1:c}, \sigma_{1:c}^2)$ summarises the model parameters and $p(c) = p(z_i = c)$. Clearly, $\sum_{c=1}^{k} p(c) = 1$.

The EM algorithm

In this section, we use intuition to introduce the expectation-maximisation (EM). If we know $\mathbb{I}_c(z_i)$, then it is easy to compute (μ_c, σ_c^2) by maximum likelihood. We repeat this for each cluster. The problem is that we have a chicken and egg situation. To know the cluster memberships, we need the parameters of the Gaussians. To know the parameters, we need the cluster memberships.

One solution is to approximate $\mathbb{I}_c(z_i)$ with our expectation of it given the data and our current estimate of the parameters θ. That is, we replace $\mathbb{I}_c(z_i)$ with

$$\xi_{ic} \triangleq \mathbb{E}\left[\mathbb{I}_c(z_i) | x_i, \theta\right]$$

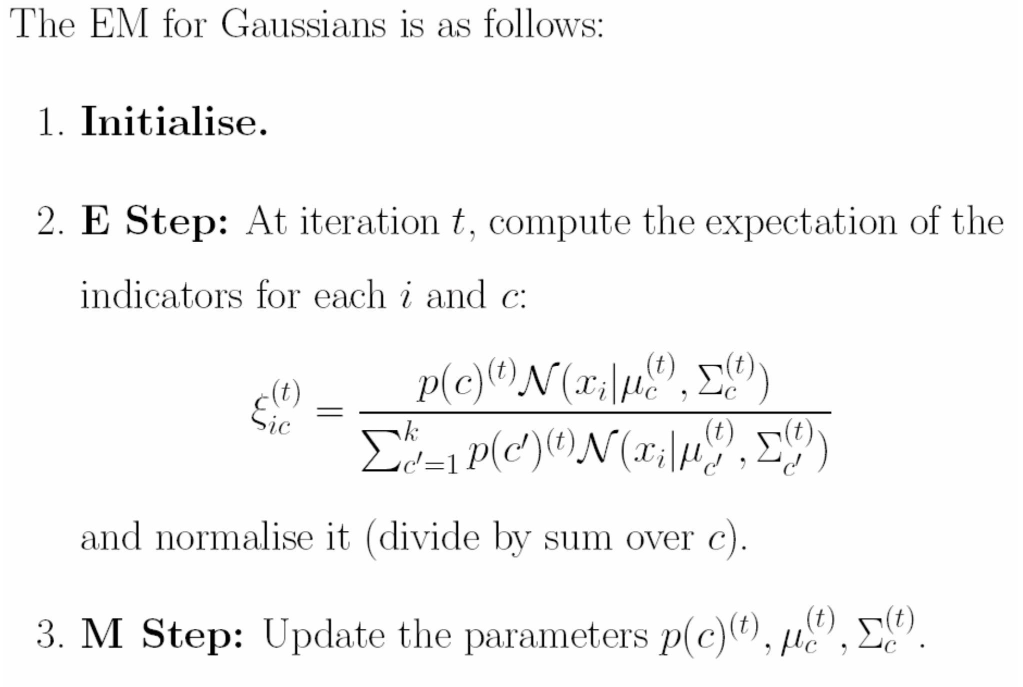
$$\xi_{ic} \triangleq \mathbb{E}\left[\mathbb{I}_c(z_i) | x_i, \theta\right] =$$

The EM algorithm

Once we know ξ_{ic}, we can compute the Gaussian mixture parameters:

$$\begin{aligned}
\mu_c &= \frac{\sum_{i=1}^n \xi_{ic} x_i}{\sum_{i=1}^n \xi_{ic}} \\
\Sigma_c &= \frac{\sum_{i=1}^n \xi_{ic}(x_i - \mu_c)(x_i - \mu_c)'}{\sum_{i=1}^n \xi_{ic}} \\
p(c) &= \frac{1}{n}\sum_{i=1}^n \xi_{ic}
\end{aligned}$$

The EM algorithm



The EM for Gaussians is as follows:

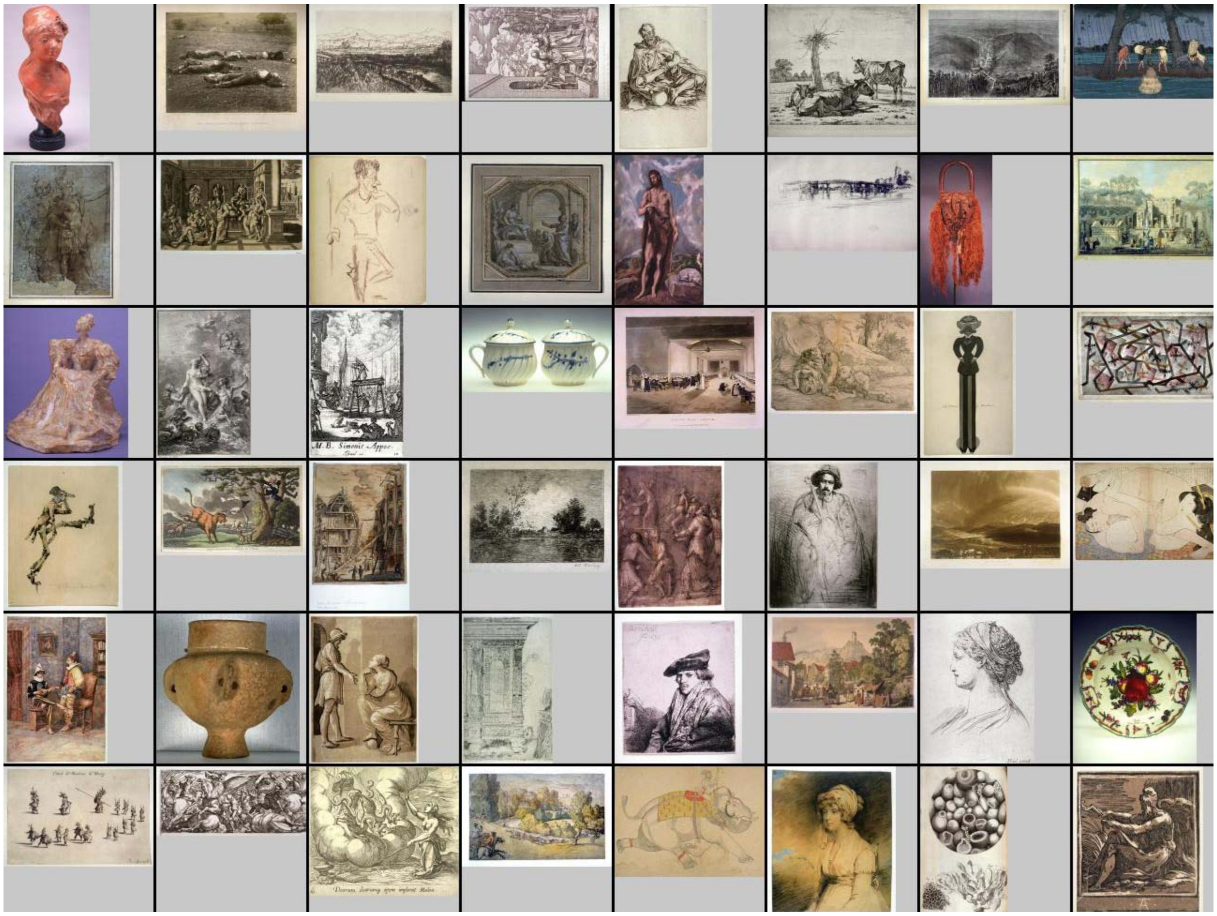
1. **Initialise.**

2. **E Step:** At iteration t, compute the expectation of the indicators for each i and c:

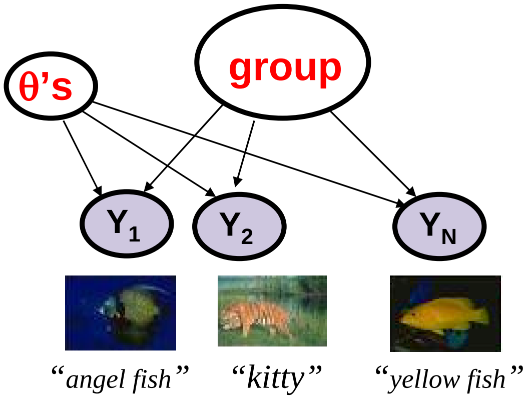
$$\xi_{ic}^{(t)} = \frac{p(c)^{(t)}\mathcal{N}(x_i|\mu_c^{(t)}, \Sigma_c^{(t)})}{\sum_{c'=1}^k p(c')^{(t)}\mathcal{N}(x_i|\mu_{c'}^{(t)}, \Sigma_{c'}^{(t)})}$$

   and normalise it (divide by sum over c).

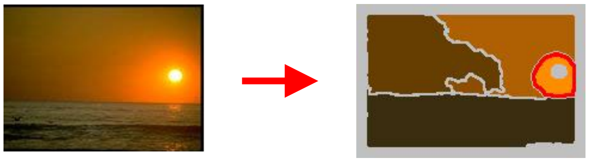
3. **M Step:** Update the parameters $p(c)^{(t)}, \mu_c^{(t)}, \Sigma_c^{(t)}$.



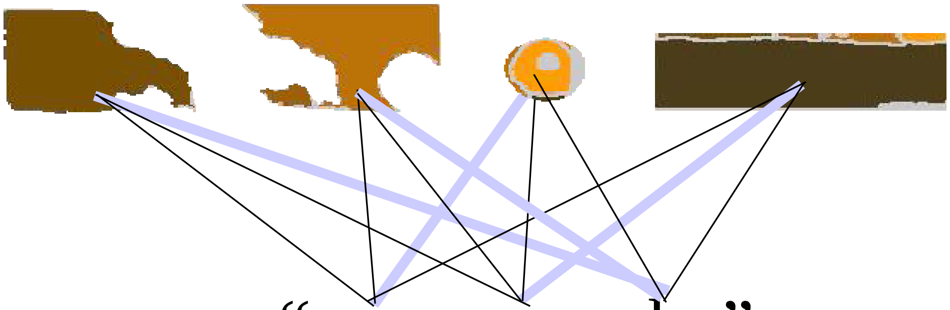
Clustering



Translation and data association



“sun sea sky”



“sun sea sky”